Open Science in Psychophysiology:

An Overview of Challenges and Emerging Solutions

Sherona Garrett-Ruffin[1], Alexandra Cowden Hindash[2], Antonia N. Kaczkurkin[3], Ryan P. Mears[4], Santiago Morales[5], Katharina Paul[6], Yuri G. Pavlov[7,8], & Andreas Keil[9]

[1] Affective Neuroscience and Mental Health Counseling, Bowling Green State University, Bowling Green, OH 43403, USA

[2] VHA Advanced Fellow in Women's Health Research, San Francisco VA Medical Center/Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, USA

[3] Department of Psychology, Vanderbilt University, Nashville, TN 37240, USA

[4] Department of Psychology, University of Florida, Gainesville, FL 32611, USA

[5] Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD 20742, USA

[6] Department of Differential Psychology and Psychological Assessment, University Hamburg, Von Melle Park 5, 20146 Hamburg, Germany

[7] Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, 72076 Tübingen, Germany

[8] Department of Psychology, Ural Federal University, Department of Psychology, Ekaterinburg, 620000, Russian Federation

[9] Department of Psychology and Center for the Study of Emotion and Attention, University of Florida, Gainesville, FL 32611, USA

Abstract

The present review is the result of a one-day workshop on open science, held at the Annual Meeting of the Society for Psychophysiological Research in Washington, DC, September 2019. The contributors represent psychophysiological researchers at different career stages and from a wide spectrum of institutions. The state of open science in psychophysiology is discussed from different perspectives, highlighting key challenges, potential benefits, and emerging solutions that are intended to facilitate open science practices. Three domains are emphasized: data sharing, preregistration, and multi-site studies. In the context of these broader domains, we present potential implementations of specific open science procedures such as data format harmonization, power analysis, data, presentation code and analysis pipeline sharing, suitable for psychophysiological research. Practical steps are discussed that may be taken to facilitate the adoption of open science practices in psychophysiology. These steps include (1) promoting broad and accessible training in the skills needed to implement open science practices, such as collaborative research and computational reproducibility initiatives, (2) establishing mechanisms that provide practical assistance in sharing of processing pipelines, presentation code, and data in an efficient way, and (3) improving the incentive structure for open science approaches. Throughout the manuscript, we provide references and links to available resources for those interested in adopting open science practices in their research.

1. Introduction

Since its formal inception during the renaissance age, organized western science has involved the sharing of theories, methods, and data within the community of scholars (Gribbin, 2002). What has once relied on letter correspondence between few experts in a given field has over time evolved into a large-scale, international industry (Lightman, 2016). At the same time, the methods used and the data obtained in fields such as psychophysiological research have become increasingly complex, reflective of technical innovation in areas such as data recording, analysis, statistical evaluation, and modeling (Kappenman & Keil, 2017). The same innovations also provide previously unheard-of opportunities for open science practices. Some of these practices have a long tradition in psychophysiology, notably sharing open stimulus sets (e.g., Bradley & Lang, 2007). There is however an emerging consensus that open science approaches provide additional, much needed benefits, when extended to data, analytical tools, and the process of study design and hypothesis testing. (Larson, 2020). Beyond addressing concerns about the replicability of published findings (Open Science Collaboration, 2015; Pashler & Harris, 2012), discussed elsewhere in this issue, open science practices may address other extant challenges in the field of psychophysiology by heightening transparency, fostering inclusivity and diversity, addressing inequalities in access to scientific resources, and ultimately helping to improve graduate and undergraduate training.

A recent study commissioned by the U.S. National Science Foundation aimed to “define reproducibility and replicability accounting for the diversity of fields in science and engineering” and to “determine if the lack of replicability and reproducibility impacts the overall health of science and engineering as well as the public’s perception of these fields” (National Academies of Sciences, 2019). Among the definitions, findings, and recommendations offered by this committee were formal definitions of reproducibility and replicability. Given their prominence in the context of open science in psychophysiology, we list some of these emerging concepts and their definitions in Table 1.

Concept	Definition/Implementation
Reproducibility & Replicability	• **(Computational) Reproducibility** is achieved when identical results are produced from archived original study data. This outcome requires access to raw data along with access to analysis code, conditions of analysis, and computing environment. • **Replicability** is achieved when the outcome of a replication study confirms or supports the original study. A replication study must match the experimental settings, measurement units, and treatments of an original study.
Open Access	**Open access** is a multifaceted construct that includes but is not limited to: • Data sharing • Data format harmonization • Workflow provenance • Pipeline sharing. **Reproducible Pipelines** are computationally reproducible analysis workflows that include code, intermediate files, electronic records of all data validation conditions and user selected signal optimizations.
Pre-Registration & Registered Reports	• **Pre-registration** refers to a practice in which researchers publicly deposit a time stamped statement regarding a planned study, minimally including a description of methods and hypotheses. • The **Registered Reports** format is a relatively standardized publication type in which a study proposal that includes theory, hypotheses, and methods (Stage 1) is peer-reviewed and published prior to data collection. The final report (Stage 2) is then accepted if consistent with the Stage 1 report, regardless of findings. This format thus fosters publication of negative findings and non-replications. For example, Registered Reports are available at this journal, and as of 2020 at the journal *Psychophysiology*.
Multi-Site Studies	Studies in which the same experimental settings, measurement units, and treatments are conducted in parallel at multiple sites such as multiple universities, multiple laboratories, etc. As such, **multi-site studies** are akin to replication studies, but are typically conducted at the same time, rather than after publication of the original study.

Table 1: Core concepts of open science practices in psychophysiology, discussed in this article.

Improving reproducibility and replicability by widely adopting open science practices may help overcome a trend in which publishers and grant agencies have incentivized research towards novel, surprising findings, often at the cost of establishing a robust premise through programmatic research (Bradley, 2017). Notably, technical innovation and increased computing power provide more options in the realm of data analysis, and also offer powerful tools for constraining hypotheses through mathematical modeling rather than expressing them in

semantic narratives (Oberauer & Lewandowsky, 2019). For example, in simulation studies, a model-driven approach of systematically testing quantitatively specified hypotheses has been shown to assist in overcoming the societal and scientific cost associated with publishing non-replicable results (Lewandowsky & Oberauer, 2020). With the rise of complex data analysis techniques available to psychophysiologists, concepts such as computational reproducibility (Table 1) have increasingly garnered attention (Keil et al., 2014). Paralleling developments in other fields of science, there is an emerging perspective that mere publication of findings from computational research is incomplete unless it is computationally reproducible. The use of proprietary, closed, and un-standardized hardware and software is unfavorable for evaluating and comparing methods and results across studies (Begley, 2013; Donoho, 2010). Furthermore, narrative and graphical communication of study results and conclusions, when offered in isolation, is unfavorable both to reproducing and to building upon prior results, if code and computing environments are not also made available (Schwab et al., 2000). These challenges have been discussed for decades, as illustrated by Buckheit and Donoho (1995) “An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”

Open science approaches are widely seen as effective in addressing these challenges. As illustrated in Figure 1, practices that enable direct replication and reproduction of experimental and analytical processes amplify the iterative benefits of hypothesis-guided but also explorative research. In the following, we identify key elements of open science, some with properties unique to the field of psychophysiology. The present report also considers the implications of open science practices for researchers at different institutions and at different career stages. Readers interested in early-career issues vis-à-vis open science are directed to the recent review by Allen & Mehler (2019). Here, we give an overview of resources and

avenues that are available to researchers and offer different perspectives regarding the potential benefits of open science practices. More specifically, we discuss challenges and perspectives for data sharing, preregistration, and multi-site studies.

2. Data and analysis pipeline sharing

This review is written at a time during which the COVID-19 pandemic is severely affecting scientific practice. This world-wide health crisis has limited many researchers' ability to collect data, travel between collaborating sites, and conduct in-person training. In this situation, the benefits of data sharing have become more apparent and have drawn attention to data sharing efforts already underway. For example, EEG/ERP researchers now have access to an open, well documented data set of high-quality EEG, recorded while the same n=40 individuals worked on six different experimental paradigms (Kappenman et al., 2020). This ERP CORE data set can be accessed at https://erpinfo.org/erp-core, together with experimental control code written in Presentation software, and analysis pipeline suggestions. Qualified researchers may also request access to the NIMH data archive at https://nda.nih.gov/, which contains harmonized, item-level data of all types, including a wide range of psychophysiological data. Researchers contributing to such sharing efforts, as well as those sharing individual experimental data in suitable repositories facilitate the benefits discussed above, including those related to fostering programmatic well-powered studies across laboratories. To maximize these benefits however, progress in the following areas is needed.

Data format standardization and harmonization. Psychophysiological data are intrinsically multivariate in nature, containing behavioral and self-report data, in combination with often several physiological measures such as heart rate, electrodermal activity, pupil diameter, respiratory rate, fMRI, EEG, MEG, and many more. Reflective of a wide range of manufacturers and industry standards, a wealth of different data formats are used for storing these

measurement modalities to disk, and for annotating them with event markers, condition labels, channel locations, etc. Although opening the primary data sometimes does not represent a major barrier for researchers, different data formats and measurement modalities tend to come with different, sometimes idiosyncratic, conventions for how event markers are recorded, how conditions are labeled, and how the data are organized within and across participants. Furthermore, psychophysiological measures differ qualitatively in their dimensionality, their digitization rate, and their spatio-temporal resolution, aggravating the unfavorable effects of variability in data organization and formatting rules between different laboratories.

Initial efforts towards harmonization have been made, aiming to standardize neuroimaging data formats, e.g. the so-called BIDS format, available for EEG, MEG, fMRI and intracranial EEG data (Gorgolewski et al., 2016; Niso et al., 2018; Pernet et al., 2019). Building on these efforts, further attempts are desirable to accommodate the needs of a wider range of scientists. Furthermore, extending harmonization efforts towards other psychophysiological measures such as electrocardiogram, electrodermal, or pupil data are needed. Standardized formats not only benefit data sharing, but are a requirement for developing widely accepted and convenient analysis pipelines that readily use a shared input format, as evident in recent developments in fMRI research (Esteban et al., 2019). Harmonization would likely benefit from adopting a scope beyond individual measures (e.g. beyond EEG/MEG) and potentially establish formats and pipelines that foster integrative or joint analysis of multi-modal data, in line with the tradition of psychophysiological research. At present, many researchers share data in the binary MATLAB “mat” format, or in other MATLAB-based formats (Delorme & Makeig, 2004). Despite their proprietary origin, these formats can be read into a variety of (non-MATLAB) analysis platforms and computing environments such as R, Python, or Julia. In addition, widespread adoption of free Python-based tools (Gorgolewski et al., 2011; Mourik et al., 2018) and meta-formats may assist in reducing the effects of remaining paywalls (e.g., for a MATLAB license, or

for commercial analysis tools). Python-based tools have also opened avenues towards harnessing the power of cloud-based, intelligent analysis pipelines that have emerged over the past decade (Zeng et al., 2020). A recent analysis of large-scale data sharing efforts in fMRI research showed that the opportunities and benefits associated with data sharing (larger sample sizes, more generalizability across different sample characteristics, financial savings) outweigh often-cited concerns (fear of being scooped, differences in data quality, usage with questionable motives), especially when effective harmonization is in place (Milham et al., 2018). Thus, it would be helpful to expand these efforts to other psychophysiological measures besides fMRI.

Visibility and searchability of shared data. Many data sharing venues exist and there is currently no widely adopted mechanism for indicating to the community where a certain shared data set can be found. As discussed above, even in situations where successful data sharing occurs, datasets are often cumbersome and esoteric, provided without data dictionaries that allow researchers to fully understand the nature of the shared data. At the time of writing, psychophysiological data are shared via local or institutional servers, via neuroscientific platforms (e.g. openneuro.org), and via unspecific repositories (e.g. the open science framework https://osf.io, databrary https://nyu.databrary.org, dryad https://datadryad.org/stash), several locations on github (e.g., https://github.com/meagmohit/EEG-Datasets), or figshare https://figshare.com). As a consequence, data may be shared but not found by interested researchers. For that reason, assigning a permanent digital object identifier is recommended, which enables searching, finding, and citing the resource (Stodden and Miguez, 2014; Stodden, et al., 2016). Finding well documented data may represent a more severe challenge for early career researchers and researchers from primarily undergraduate institutions, compared to established PIs with extensive professional networks who typically have more opportunity for exchange with other researchers at grant review panels, conferences, and through the journal review and editing process.

One straightforward way for making data visible is to connect a publication with its underlying data in a repository, which increasingly occurs at the stage of a preprint publication (Cragin et al., 2010). Preprint servers such as PsyArXiv and BioRxiv are widely used and well suited for psychophysiological research. In addition, several journals offer repositories in which data can be shared and linked to the respective publication. Although this implies that data files need to be organized in a fashion that allows sharing already during the manuscript writing phase, an increasing number of authors now opt for the data being made public after publication (e.g. after an embargo period) or for data being made public without a corresponding publication. In this regard, ethical and intellectual property aspects gain additional significance (Carroll, 2015). For example, embargoes may be implemented in order to protect early career researchers, or researchers from laboratories with limited resources, from their data being used, perhaps more rapidly, by those with more abundant resources.

Variability of experimental procedures. Paradigm sharing is made difficult by the wide array of software solutions used for stimulus presentation and response registration. Paralleling data formats, there are lab-specific idiosyncrasies in terms of how stimulus control software interacts with the recording environment and in terms of how event markers are sent and stored. Event markers may be stored in the data file as a mere time stamp, to be matched with condition names in an external log file or dedicated marker file, or detailed condition codes may be stored as part of the data or header file. Although some of these sources of variability can be addressed by extant standard formats available for neuroimaging measures, such as BIDS (Gorgolewski et al., 2016; Niso et al., 2018; Pernet et al., 2019), there are several remaining barriers that impede the successful sharing of paradigms and data. The authors identified the following practical steps towards overcoming these barriers, some of which are already being implemented:

First, broad sharing of paradigms and experimental control code, despite diversity in coding and formatting, heightens the probability that a researcher will find a given paradigm in their preferred platform, such as Presentation, PsychoPy, Psychtoolbox, E-Prime, etc. In addition, multi-site, coordinated studies (see section below) assist in identifying the amount of convergence/divergence between standard paradigms (e.g., an arrow flanker task, a picture viewing task), which ultimately enables forming a library of standard paradigms for multiple presentation/stimulation platforms without asking researchers to adopt one common standard presentation software—widely seen as an unreasonable and unpractical approach. Researchers may also want to share paradigms together with the resulting data to allow comparison of outcomes with standard paradigms across different laboratories, or to compare their own data with widely accepted gold-standard data (e.g., Kappenman et al., 2020). Likewise, calibration scripts that display simple stimuli at known timing and spatial locations may well be shared among labs to establish convergence/divergence of timing accuracy and psychophysiological outcome measures with different stimulus hardware and recording setups present in different laboratories.

Second, precise documentation of the presentation setup (monitor, recording setup, stimulation hardware and software) is encouraged by extant guideline papers (e.g., Keil et al., 2014). Such precise reporting in published papers enables replication of setups, particularly relevant for researchers about to establish their own laboratories. Registered reports, which tend to provide greater detail regarding stimulus presentation and data analysis, are therefore particularly helpful in the context of paradigm sharing.

Variability of analysis workflow procedures*.* Psychophysiological data are composed of multivariate time series. A substantial range of algorithms exist for cleaning and analyzing these signals. Although the diversity in algorithms is greatly beneficial for addressing a wide variety of problems and questions, the number of algorithms and the lack of gold standard

methods pose a significant challenge to reproducibility. During preprocessing and data analysis, researchers make choices that are often simultaneously justifiable, motivated, and arbitrary (Simonsohn et al., 2019). For instance, in a recent report from the Neuroimaging Analysis Replication and Prediction Study, 70 independent research teams analyzed the same fMRI dataset and no two teams used the same workflow pipelines (Botvinik-Nezer et al., 2020). Although few such studies exists for other psychophysiological data (Miltner et al., 1994; Drisdelle et al., 2017; Sandre et al., 2020) the flexibility and diversity in preprocessing and data analysis pipelines is comparable across psychophysiology methods (e.g., EEG, MEG, and fMRI). This emphasizes the need for a detailed description of the methods used in publications (Keil et al., 2014). Another, and perhaps more desirable approach may be to share the full algorithmic pipeline in addition to the data, to allow other researchers to perform an in-depth analysis of the methods, reproduce the analysis, and/or apply them to their own data. However, exacerbating the problem, these preprocessing and data analytic algorithms are often implemented in different programming environments and vary in their availability. For example, some algorithms use commercial or precompiled user interfaces, and many are specific to a psychophysiological measure or to a given piece of hardware. Thus, sharing analytical pipelines does not always have the desired outcome of enabling other researchers to reproduce the analysis, because they may not have access to the software needed, may not know how to use it, and may not have the required hardware. Increasingly, free versions of previously restricted algorithms exist, in various computing environments. Furthermore, open source analysis toolboxes for EEG/MEG analysis are increasingly used in the field, many with plugins for additional psychophysiological measures. These include, but are not limited to, the following toolboxes: Brainstorm (Tadel et al., 2011), EEGLAB (Delorme & Makeig, 2004), emegs (Peyk et al., 2011), FieldTrip (Oostenveld et al., 2010), MNE/MNE-Python (Gramfort et al., 2013), and SPM (Litvak et al., 2011).

Even when researchers are able to access the preprocessing and data analysis pipelines shared by others, the majority of analytical methods routinely used in psychophysiology require user intervention. If such methods are to be successfully shared, they would therefore need to be accompanied by specific instructions to ensure exact replication. These instructions are often implicit or depend on the user's expert knowledge or extensive training (see e.g. Miltner et al., 1994). Depending on the subjective judgment of researchers, this reliance on expert knowledge may present a further obstacle to replication, which can be especially problematic for large, multi-site studies in which one expert cannot analyze all the data. Recent open source efforts have developed preprocessing and data analytic pipelines that overcome several of the challenges listed above by automating steps that require user input (e.g., ADJUST, Mognon et al., 2011; ICLabel, Pion-Tonachini et al., 2019; FASTER, Nolan et al., 2010; Adjusted ADJUST, Leach et al., 2020) or providing fully automated pipelines that can be implemented by other research groups relatively easily. Some of the existing pipelines include the PREP pipeline (Bigdely-Shamlo et al., 2015), HAPPE & BEAPP pipeline (Gabard-Durnam et al., 2018; Levin et al., 2018), MADE pipeline (Debnath et al., 2020), EPOS pipeline (Rodrigues et al., 2020), and CTAP toolbox (Cowley et al., 2017). Some pipelines automatize the estimation and application of parameter settings previously set by the user (Engemann & Gramfort, 2015; Jas et al., 2017). Several of these pipelines allow users to set specific parameters that may improve pipeline performance on a particular dataset. These parameters can then be reported, allowing others to replicate the results obtained.

The sharing of analytical pipelines ultimately relies on collaboration and the exchange of code that others can understand and adapt. This requires excellent documentation, including examples that allow others to more easily comprehend the algorithms' functions (Eglen et al., 2017). It also requires writing clean and well-commented code (Cohen, 2017). The authors recognize that advancing computational reproducibility through pipeline sharing requires new

efforts in training at the graduate and undergraduate level, providing new researchers with powerful computing and documentation skills. In addition, a growing number of universities have employed Research Software Engineers, who support the development and maintenance of sustainable and replicable research computing environments (Cohen et al., 2020). Where intra-institutional assistance is not feasible, various online communities are open and accessible. Organizations and programs that promote open development and sharing of code range between prudently structured training to unstructured but extensive exchange of information.

New programs such as the Code Refinery initiative (http://coderefinery.org), available in Nordic and Baltic countries also provide support along with storage and curation services and training for researchers interested in sharing clean, reproducible, computer code. ROpenSci (https://ropensci.org) utilizes a framework for the review and maintenance of open source scientific code (Ram et al, 2018). Although the ROpenSci community centers on code written in the R language, there are parallel efforts to adapt these practices toward code review based on other programming languages. The Carpentries offer training in use and development of scientific code in R and Python as well as pedagogical training to teach and facilitate open practices (Shade and Teal, 2015; Wilson, et al., 2017; Wilson, 2019). Many organizations include focus closely on the side of specific psychophysiology disciplines: BrainHack (https://brainhack.org), ICNF (https://training.incf.org), NeuroStars (https://neurostars.org), ReproNim https://www.repronim.org), Neurodata without borders (https://www.nwb.org), NeuroVault (https://neurovault.org).

Psychophysiological research is likely to benefit from utilizing novel ways for sharing and illustrating code through applications that provide interactive documents (e.g., live scripts in MATLAB), integrate multiple programming languages (e.g., Jupyter Notebooks, Rule, et al., 2019), and even permit video streaming or recordings of the actual data analysis process (e.g., YouTube or Twitch). In addition to providing a more transparent data analysis process, sharing data analysis pipelines and making these pipelines more accessible by leveraging these new

tools could serve as a training resource for others, encourage good coding habits, and promote reproducibility in psychophysiology.

Comparison of analysis pipelines. Future research will systematically quantify the convergence and difference of similar analytical procedures (e.g., different types of wavelet analysis, or different algorithms for blink interpolation in pupil data), as well as examining the impact of other decisions during data analysis (see next section). So-called multiverse studies (Steegen et al., 2016) may assist in this process. One obstacle towards this goal is that the criteria for evaluating and comparing pipelines are currently unclear. Desirable characteristics would be pipelines that a) maximize the usage of available data (i.e. do not discard excessive amounts as artifact), b) provide the best signal-to-noise ratio, c) yield more reliable measures, and d) follow a "Glass Box" philosophy (i.e., automated, but transparent). Importantly, some of these characteristics may differ by study characteristics such as participant population, hardware, experimental procedures, or measures of interest. As such, examining which pipelines or which algorithms within different pipelines perform best under which circumstances represents an important first step towards developing gold standard data analysis pipelines. Indices for quantifying data quality in a unitless fashion are useful steps towards this goal. Psychophysiology has a long tradition of reporting signal-to-noise ratios for dependent variables (Regan, 1989), and more recently developed indices of data quality in ERP research also hold promise for objectively assessing data quality (Junghöfer et al., 2000; Luck et al., 2020). Widely using and reporting such measures will aid transparency and reproducibility.

Although the characteristics described above are crucial for maximizing data usage and obtaining reliable measurements in a transparent and consistent manner, they do not address concerns about the construct validity of the measures obtained. One promising way to start quantifying the impact of different preprocessing and data analysis decisions are specification-curve (Simonsohn et al., 2019) and multiverse-analysis (Steegen et al., 2016) approaches.

Rather than presenting one analysis pipeline, these studies involve performing all reasonable analytic steps using reasonable specifications. Such an approach can help determine the impact that different (and often arbitrary) choices in data preprocessing and data analysis have on the results and conclusions. Thus, specification-curve and multiverse-analysis approaches may provide novel insights into the impact that analysis pipelines have on the relations between psychophysiological measures and the theoretical constructs or outcomes of interest.

3. Preregistration

Central goals of preregistration are to increase study transparency and to foster systematic and programmatic research. Preregistration encourages a researcher to consider and publicly state multiple facets of the project prior to data collection and analysis. For preregistration of psychophysiological studies, the Open Science Foundation (OSF) and University of Pennsylvania and Wharton School Credibility Lab's AsPredicted.org offer the most compatible formats. Preregistration involves: (1) identifying study contributors; (2) detailing hypotheses; (3) detailing the research design and sampling plans – including a sample size rationale; (4) specifying variables; (5) detailing data processing and analysis plans – including data exclusion criteria; (6) Other important information such as exploratory data considerations, potential contributor changes, etc. The open science framework contains examples of ERP and fMRI preregistrations (see e.g., Paul et al., 2020).

Preregistration can take different forms, from registration of study goals on a suitable online platform, to a two-stage registered report formally overseen by one of many journals who offer this format (see e.g., Keil et al., 2020). It has often been noted that despite increasing transparency and accountability, preregistration practices are not a panacea for addressing all problems that have led to low replicability in biomedical and behavioral research (Chambers, 2019b). In the case of psychophysiological research, preregistration does not overcome poor

statistical practices, lack of a systematic research program, or limitations of power analyses, nor does it address inadequate theory and lack of quantitative models (Szollosi et al., 2020). However, use of preregistration can help researchers and reviewers differentiate what aspects of a study were planned and what aspects were exploratory. Specifically, preregistration functions to address issues such as underreporting null findings and questionable scientific practices such as hypothesizing after the results are known (i.e, HARKing, see Figure 1) and flexible data inclusion/exclusion (i.e., cherry-picking) decision making (Chambers, 2019a). The clear differentiation between a-priori hypotheses and exploratory analysis allows more rigorous hypothesis testing as well as more transparent exploratory research. Furthermore, students and early career scientists may benefit from a pre-registration, or stage 1 Registered Report in different ways. For example, preregistration allows for a published record of a researcher's contribution to a study, even if study completion takes a long time, or if the researcher leaves a laboratory before data collection has been completed or the data have been published.

Many of the advantages of preregistration are particularly relevant to psychophysiological research: Pre-specifying recording and data analysis pipelines, data reduction steps, and the composition of dependent variables assists in reducing researcher degrees of freedom (Wicherts et al., 2016). Publishing the full code and processing steps used facilitates computational reproducibility, while also enabling the scientific community to catch errors in the code, or clarify any misconceptions as to how it is used. Thus, these aspects of pre-registration address not only questionable practices but also assist in preventing and managing honest mistakes and oversights, which prompt paper retractions. When selecting and aggregating high dimensional psychophysiological data into low dimensional variables for statistical analysis, such researcher degrees of freedom are obvious. Harmonized data formats (see discussion above) and widely applicable analysis pipelines (as discussed below) are currently emerging and may increasingly assist with pre-registration in the future. Finally, efforts

underway in many laboratories and some journals (Marcus, 2016) to formalize the content of Methods sections in algorithmic and tabular form, rather than as a narrative, will ultimately assist in matching pre-registered steps with the steps actually performed, thus fostering reproducibility. A recent standardization initiative for EEG data (Styles, Kovic, Ke, Šoškićis) is described on the Open Science framework: https://doi.org/10.17605/OSF.IO/PVRN6.

Power analysis. Reviewers and authors alike are increasingly aware of the fact that sample sizes should be justified. A widely encouraged means to accomplish this goal is traditional power analysis, which is grounded in null-hypothesis testing (Button et al., 2013). Despite these efforts, it can be observed anecdotally that broad statements regarding sample sizes in the absence of quantitative analyses are abundant. Science twitter, reviewers' comments, and conference conversations often include notions to the effect that "20 is not enough" irrespective of effect size or paradigm. The present authors consider it desirable that sample sizes be based on appropriate, quantifiable methods, and that anecdotal or intuition-based judgments be minimized. However, estimating the required sample size is not trivial. The group observed several challenges with respect to statistical power in psychophysiology.

The first is that small sample sizes do not equal low power: Many studies reporting some of the highly replicable standard effects in psychophysiology were based on small samples (e.g., in EEG/ERP research the P300 effect, the LPP effects, P1 spatial attention effects, alpha blocking). These effects have been shown to be robust and have been replicated hundreds of times in studies where the technical execution was done correctly and where the signal-to-noise ratio of the dependent variable was acceptable. This highlights the important role of two factors: Effect size and data quality (Clayson et al., 2013; Thigpen et al., 2017). Many researchers are interested in smaller effects than those mentioned above, often because they are interested in additional variables, e.g. they may ask: how is the P1 spatial attention effect modulated by threat? These researchers will not be able to base their sample size choices on very strong

effects such as the P1 spatial attention main effect. Instead, they will have to use some form of power analysis or simulation study to estimate a more realistic (larger) sample size needed for their study (Gibney et al., 2020). A detailed discussion of simulation studies is outside of the scope of the paper but pertinent examples have recently been published (Boudewyn et al., 2018).

Second, given the multivariate nature of psychophysiological research, power analyses for within-participants (repeated measures) designs are highly sensitive to inter-variable correlations. These correlations yield dramatically different required sample sizes depending on the strength of the inter-variable correlations expected. These are however very rarely reported in the published literature, and they may vary depending on the equipment used, the within and cross-trial timing of the study, and the noise level of the past or expected data (P. E. Clayson et al., 2013).

This leads to the third challenge: Statistical power and the required sample size to detect an effect are both influenced by data quality. As a consequence it is desirable for researchers to know the trial-by-trial variability and other low-level parameters of the data. These are however not always available when researchers use turn-key systems that output only processed variables such as for example theta-beta ratios in EEG feedback research. Finally, the concept of statistical power is closely tied to null hypothesis significance testing (NHST), and as such part of a larger discussion in which problems of NHST have prompted efforts towards alternate statistical methods, including Bayesian approaches. For example, Bayesian rules for stopping data collection are now available (Schönbrodt & Wagenmakers, 2018) which are not rooted in the paradigm of NHST and guide researchers into using sample sizes that provide sufficient evidence for or against a given hypothesis.

In order to address challenges related to power, several practical steps can be taken. First, reporting the rationale for sample size decisions has increasingly been encouraged by many journals and grant agencies. Given the above considerations, this practice is expected to have a positive effect on replicability and transparency. Second, sample sizes should reflect the expected effects while also modeling the properties of the psychophysiological measure of interest (e.g. the signal-to-noise ratio) as well as the analytical plan (including artifact control, and averaging procedures, etc). Traditional power analysis for within-participants (repeated measures) designs in software such as G*power requires exact knowledge of inter-variable correlations (Guo et al., 2013). Thus, if researchers in the field habitually reported the intercorrelations of the dependent variables, or made available the data matrix, then the realism and quality of power analyses could be dramatically enhanced.

Finally, traditional power analysis may not capture aspects of contemporary statistical approaches, e.g. those in which a computational model is fitted to the data, or those involving machine learning. In those and many other cases, it is recommended that power and sample size be calculated based on suitable simulations. These simulations may take into account the covariance structure, signal-to-noise, and temporal stability of the data contributing to measuring dependent variables. For example, as compared to the widely used G*Power, several packages exist that are capable of estimating sample size for a study with fully within-subject design, common in psychophysiology, (e.g., 2 x 2 x 2: Condition x Time window x Channel interaction). R has several power packages available also as Shiny apps (for example PANGEA https://jakewestfall.shinyapps.io/pangea/, Superpower http://arcaldwell49.github.io/SuperpowerBook), and MOREpower (https://wiki.usask.ca/display/MorePowerCalculatorV6/Home), which is standalone software.

4. Multi-Site Studies

Multi-site studies, in which the same research is conducted at different sites, are desirable because they enable researchers to increase statistical power by increasing the total sample size of the study, promote transparent practices, facilitate communication between researchers, and foster quality control. Pooling data from multiple sites is critical in studies of rare disorders and other populations that are difficult to recruit from (Smith et al., 2020; Swerdlow et al., 2007). Replication of a given effect across different laboratories tends to increase confidence in the robustness of that effect, especially in the case of surprising or counter-intuitive effects (Bekhtereva et al., 2018). Similarly, multi-site studies encourage careful research practices, increase generalizability of the findings, help to avoid mistakes likely to be overlooked by a single researcher, and distribute the work between participating researchers, often reducing the overall workload (Johnson et al., 2009). For example, simultaneously running the same experimental paradigm in two EEG laboratories using different-brand EEG systems assists in expanding the sample size of the full study, and it establishes generalizability across hardware platforms and specific populations (Bekhtereva et al., 2018). Despite these potential benefits, multi-site studies are relatively rare in psychophysiology. Nevertheless, some examples exist (Nave et al., 2018; Nieuwland et al., 2018; Pavlov et al., 2020; Whiteford et al., 2019). Additionally, multi-site collaborative studies also pose a number of challenges as discussed next.

Funding. Despite the increasing importance of collaborative research, most funding agencies do not have programs for supporting research at multiple institutions spread over the world. A recent example is shown in the reluctance of US funding agencies to support the Psychological Science Accelerator (Moshontz et al., 2018) despite its longstanding dedication to the promotion of reproducible, inclusive, and generalizable research. As a consequence, most multi-site studies are forced to limit their overall costs by refraining from using complex setups

and expensive equipment. As a partial solution, local foundations may provide funding to core sites for supporting their infrastructure. This arrangement may make distribution of funds across sites challenging, especially if political barriers are in place, such as the embargo of Iran, Cuba, and other countries by the United States of America.

Early career researchers tend to be particularly responsive to opportunities for participating in large-scale collaborative projects (Allen & Mehler, 2019). However, the challenges of conducting a multi-site study may disproportionately dissuade early career investigators who are not yet established in the field, have fewer available collaborators, and may not yet have the academic caliber to convince funders of their ability to execute a multi-site study. Similar constraints may apply to researchers at primary undergraduate institutions, or investigators in laboratories that are less funded than some of their peers. Given the growing emphasis on obtaining large sample sizes across many disciplines, multi-site studies as well as studies based on openly shared data may become increasingly desirable for journal reviewers and editors. Scientific societies and funding agencies may positively impact these challenges by providing training opportunities and specific financial and infrastructure resources to those interested in pursuing multi-site studies. Furthermore, creating positions for research software engineers, as mentioned previously, may represent another helpful step towards integrating paradigms and data across collaborating sites.

Coordination. Coordination of a multi-site study involves multiple challenges. Researchers need to identify collaborators who are willing and able to invest their time and resources, choose site locations to ensure a relatively representative sample, convince funding sources of the feasibility of the project, and coordinate ethics policies that vary by institution. Additionally, researchers need to navigate complex subcontracts, delegate funding and responsibilities between sites, coordinate communication between sites at all steps in the research process, plan for setbacks in costs or recruitment that may vary by site, and ensure fair

authorship credit for all involved. Issues related to institutional regulations are especially challenging for international collaborations, as policies differ widely by country (Arellano et al., 2018; Chassang, 2017; Dove, 2018). Effective coordination of multi-site studies is especially challenging without dedicated personnel and administrative resources. Thus, a multi-site study is a risky undertaking for early career researchers who may have less funding than established senior researchers and less freedom to take chances as they seek tenure.

Cross-laboratory harmonization. Between-sites differences in hardware (e.g., EEG amplifiers, eye trackers, MRI pulse sequences, etc.) add unwanted variability to the data that may prevent pooling data collected in different labs. Many of the challenges discussed under "Data sharing", above, apply here as well. Data harmonization between labs to control for equipment manufacturer, location, and cohort differences remains a major challenge of setting up a multi-site study. In fMRI research, multiple studies have successfully reduced or eliminated site effects (Yamashita et al., 2019; Yu et al., 2018). There is a critical need for other domains of psychophysiological research to find a solution for this problem as well.

One potential solution, the "travelling subject" approach, has been useful for testing the efficacy of harmonization in fMRI collaborations (Sutton et al., 2008). It involves scanning the same participant at multiple sites and also the same number of times at the original site (e.g., if there are five sites, then 6 recordings are compared: 2 from the first site and 4 from the other ones). For example, in Whiteford et al. (2019), the first author had her EEG recorded at each participating site. Five of the six sites used the same type of amplifier. In this study, within-site reliability of the EEG recordings was not significantly different from between-site reliability. Ultimately, the development of an affordable artificial participant that can be used for calibration and cross-validation would be desirable, such as phantom heads used in fMRI and MEG research. Another way to account for differences in hardware is to pool not raw data but derivatives such as standardized effect sizes and normalized values of dependent variables

(e.g. peak-scored skin conductance converted to z-scores, independent EEG/MEG components instead of single channel EEG/MEG). This approach makes harmonization easier to achieve but limits the diversity of potential analyses.

Another challenge to a multi-site approach is the difficulty of establishing consistent quality standards across participating laboratories. For example, exact locations in EEG montages may differ between laboratories even when using a system of the same brand, with the same number of channels, because researchers may have configured the channels differently. Similar issues have been noted with respect to MR sequences and filter settings in recordings of autonomic physiology or MEG. It is often neglected that the technical expertise varies across different laboratories, representing a challenge for quality control in multi-laboratory, collaborative studies. To address these problems, lab visits among collaborators may be helpful, as required in clinical trial protocols. At the same time, there is an absence of cross-laboratory gold standard indices for establishing the same recording and data quality (Farzan et al., 2017). Likewise, there are no widely established methods for achieving cross-validation of findings, and available guidelines for how to achieve common signal quality on different recording systems are not yet widely adopted in the field. As discussed above, researchers may compute signal-to-noise ratios (e.g., Regan, 1989), as well as quality indices based on waveform and trial variability, which are mathematically unchallenging, unitless, and applicable across measurement modalities (Junghöfer et al., 2000; Luck et al., 2020).

Coordination of analytical strategies. A final challenge with multi-site studies arises at the level of data processing and analysis. Multi-site studies as well as analyses of large shared data sets require scalability of analysis pipelines from few participants to hundreds of participants. Not all methods are scalable, highlighting the need to consider this point at the time of study planning. Additionally, there is the issue of what level of processing shared data will undergo. Some multi-site collaborations share aggregated data at the level of group means or

effect sizes (e.g., for meta-analyses), others may choose to share data that has undergone basic processing locally, while yet others may want to share raw unprocessed data which will then be processed by one site only (e.g., mega-analyses). For collaborations in which partially processed or raw individual data are shared, de-identification of shared data is critical to remove any personally identifiable information that could violate participant confidentiality (Moctezuma & Molinas, 2020). Likewise, transferring data between sites requires adequate encryption and security measures. In addition, the pooled data will need to be organized in such a way to facilitate processing (see the discussion on BIDS formatting above). Furthermore, as mentioned earlier, preregistration of the analytic plan has been described as a successful strategy (Chambers, 2019b; Nosek et al., 2018; Wagenmakers et al., 2012). Here, multi-site studies face the additional obstacle of having to reach a consensus regarding analytical strategies across multiple investigators with potentially diverse views.

5. Solutions

The authors identified several practical steps that may be taken by individual researchers to foster open science practices in psychophysiology.

Computing and reporting indices of data quality and reliability. Although there have been efforts towards establishing an objective, system-independent index of data quality, applicable to shared EEG/ERP, these efforts are not yet widely adopted. Indices of data quality are readily computed and widely available (Junghofer et al., 2000; Luck et al., 2020; Regan, 1989). In a similar vein, calculating and reporting metrics of reliability and internal consistency (Clayson & Miller, 2017a, 2017b; Thigpen et al., 2017) contributes to harmonization and fosters sharing and comparing open science data. Other efforts, such as showing standard errors of physiological time series and routinely reporting signal-to-noise ratios will serve a similar purpose. It was noted that leveraging multi-level statistical models, increasingly used in psychophysiology,

assists in explicitly modeling and thus quantifying systematic variance between laboratories that attempt to perform the same study.

Database for open calls for collaboration in psychophysiology. It was observed that there is a substantial appetite among researchers at different career stages for engaging in open-science multi-laboratory research, but the communication of collaborative opportunities is perceived as lacking. Establishing a platform for facilitation of collaborative studies represents a task better suited for scientific societies than for individual researchers.

Funding database for multi-site studies. Scarce funding opportunities for multi-site (especially) international studies limit the ability of researchers to engage in robust, multi-site studies. It would be desirable to develop a database of funding opportunities for (1) promoting open science initiatives (2) national grants with open science, collaborative research focus (3) international grants to support multi-site studies.

Assistance for transparent coding and sharing of processing pipelines. With programming languages and platforms in constant flux, researchers who focus on conceptual, applied, or clinical research may not be in a position to share their processing pipelines in the best way for others to find, understand, and reproduce. Major steps in this regard would involve establishing training programs, online resources, and mechanisms that provide practical assistance for researchers who seek to share their code in an efficient way.

Training. As a final point, the authors observe that training in the skills needed to implement the recommendations above is not yet widely available. Goals for training in the field of psychophysiology include training in what is under the hood of widely used programs for data reduction, analysis, and statistical evaluation. It also includes training in the mathematical and biophysical foundations that enable linking concepts such as signal-to-noise to methodological constraints related to measurement, and eventually enable a researcher to perform simulation-

based power analysis. Such training would be most effective if it were deployed in a broad and accessible, open, fashion.

References

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), e3000246. https://doi.org/10.1371/journal.pbio.3000246

Arellano, A. M., Dai, W., Wang, S., Jiang, X., & Ohno-Machado, L. (2018). Privacy Policy and Technology in Biomedical Data Science. *Annual Review of Biomedical Data Science*, *1*(1), 115–129. https://doi.org/10.1146/annurev-biodatasci-080917-013416

Begley, C. G. (2013). Reproducibility: Six red flags for suspect work. *Nature*, *497*(7450), 433–434. https://doi.org/10.1038/497433a

Bekhtereva, V., Pritschmann, R., Keil, A., & Müller, M. M. (2018). The neural signature of extracting emotional content from rapid visual streams at multiple presentation rates: A cross-laboratory study. *Psychophysiology*, *55*(12), e13222. https://doi.org/10.1111/psyp.13222

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. https://doi.org/10.3389/fninf.2015.00016

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(6), e13049. https://doi.org/10.1111/psyp.13049

Bradley, M. M. (2017). The science pendulum: From programmatic to incremental—and back? *Psychophysiology*, *54*(1), 6–11. https://doi.org/10.1111/psyp.12608

Bradley, M. M. L., P. J. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 29–46). Oxford University Press.

Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and Reproducible Research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and Statistics* (Vol. 103, pp. 55–81). Springer New York. https://doi.org/10.1007/978-1-4612-2544-7_5

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carroll, M. W. (2015). Sharing Research Data and Intellectual Property Law: A Primer. *PLOS Biology*, *13*(8), e1002235. https://doi.org/10.1371/journal.pbio.1002235

Chambers, C. (2019a). *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton University Press.

Chambers, C. (2019b). What's next for Registered Reports? *Nature*, *573*(7773), 187–189. https://doi.org/10.1038/d41586-019-02674-6

Chassang, G. (2017). The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience*, *11*. https://doi.org/10.3332/ecancer.2017.709

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology*, *50*(2), 174–186. https://doi.org/10.1111/psyp.12001

Clayson, Peter E., & Miller, G. A. (2017a). ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related brain potentials. *International Journal of Psychophysiology*, *111*, 68–79. https://doi.org/10.1016/j.ijpsycho.2016.10.012

Clayson, Peter E., & Miller, G. A. (2017b). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology*, *111*, 57–67. https://doi.org/10.1016/j.ijpsycho.2016.09.005

Cohen, J., Katz, D. S., Barker, M., Hong, N. P. C., Haines, R., & Jay, C. (2020). The Four Pillars of Research Software Engineering. *IEEE Software*, 0–0. https://doi.org/10.1109/MS.2020.2973362

Cowley, B. U., Korpela, J., & Torniainen, J. (2017). Computational testing for automated preprocessing: A Matlab toolbox to enable large scale electroencephalography data processing. *PeerJ Computer Science*, *3*, e108. https://doi.org/10.7717/peerj-cs.108

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165

Debnath, R., Buzzell, G. A., Morales, S., Bowers, M. E., Leach, S. C., & Fox, N. A. (2020). The Maryland analysis of developmental EEG (MADE) pipeline. *Psychophysiology*, *57*(6), e13580. https://doi.org/10.1111/psyp.13580

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/J.Jneumeth.2003.10.009

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028

Dove, E. S. (2018). The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era. *The Journal of Law, Medicine & Ethics*, *46*(4), 1013–1030. https://doi.org/10.1177/1073110518822003

Drisdelle, B. L., Aubin, S., & Jolicoeur, P. (2017). Dealing with ocular artifacts on lateralized ERPs in studies of visual-spatial attention and memory: ICA correction versus epoch rejection. *Psychophysiology*, *54*(1), 83–99. https://doi.org/10.1111/psyp.12675

Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P., Silver, R. A., Davison, A. P., Lanyon, L., Abrams, M., Wachtler, T., Willshaw, D. J., Pouzat, C., & Poline, J.-B. (2017). Toward standard practices for sharing computer code and programs in neuroscience. *Nature Neuroscience*, *20*(6), 770–773. https://doi.org/10.1038/nn.4550

Engemann, D. A., & Gramfort, A. (2015). Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage*, *108*, 328–342. https://doi.org/10.1016/j.neuroimage.2014.12.040

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Farzan, F., Atluri, S., Frehlich, M., Dhami, P., Kleffner, K., Price, R., Lam, R. W., Frey, B. N., Milev, R., Ravindran, A., McAndrews, M. P., Wong, W., Blumberger, D., Daskalakis, Z. J., Vila-Rodriguez, F., Alonso, E., Brenner, C. A., Liotti, M., Dharsee, M., … Kennedy, S. H. (2017). Standardization of electroencephalography for multi-site, multi-platform and multi-investigator studies: Insights from the canadian biomarker integration network in depression. *Scientific Reports*, *7*(1), 7473. https://doi.org/10.1038/s41598-017-07613-x

Gabard-Durnam, L. J., Mendez Leal, A. S., Wilkinson, C. L., & Levin, A. R. (2018). The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): Standardized Processing Software for Developmental and High-Artifact Data. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00097

Gibney, K. D., Kypriotakis, G., Cinciripini, P. M., Robinson, J. D., Minnix, J. A., & Versace, F. (2020). Estimating statistical power for event-related potential studies using the late positive potential. *Psychophysiology*, *57*(2), e13482. https://doi.org/10.1111/psyp.13482

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00267

Gribbin, J. (2002). *Science: A History 1543-2001*. Penguin Books.

Guo, Y., Logan, H. L., Glueck, D. H., & Muller, K. E. (2013). Selecting a sample size for studies with repeated measures. *BMC Medical Research Methodology*, *13*(1), 100. https://doi.org/10.1186/1471-2288-13-100

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417–429. https://doi.org/10.1016/j.neuroimage.2017.06.030

Johnson, E. I., Grondin, O., Barrault, M., Faytout, M., Helbig, S., Husky, M., Granholm, E. L., Loh, C., Nadeau, L., Wittchen, H.-U., & Swendsen, J. (2009). Computerized ambulatory

monitoring in psychiatry: A multi-site collaborative study of acceptability, compliance, and reactivity. *International Journal of Methods in Psychiatric Research*, *18*(1), 48–57. https://doi.org/10.1002/mpr.276

Junghofer, M., Elbert, T., Tucker, D. M., & Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology*, *37*(4), 523–532.

Kappenman, E., Farrens, J., Zhang, W., Stewart, A. X., & Luck, S. J. (2020). *ERP CORE: An Open Resource for Human Event-Related Potential Research* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/4azqm

Kappenman, E. S., & Keil, A. (2017). Introduction to the special issue on recentering science: Replication, robustness, and reproducibility in psychophysiology. *Psychophysiology*, *54*(1), 3–5. https://doi.org/10.1111/psyp.12787

Keil, A., Debener, S., Gratton, G., Junghofer, M., Kappenman, E. S., Luck, S. J., Luu, P., Miller, G. A., & Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21. https://doi.org/10.1111/psyp.12147

Keil, Andreas, Gatzke-Kopp, L. M., Horváth, J., Jennings, J. R., & Fabiani, M. (n.d.). A registered report format for Psychophysiology. *Psychophysiology*, *n/a*(n/a), e13663. https://doi.org/10.1111/psyp.13663

Larson, M. J. (2020). Improving the Rigor and Replicability of Applied Psychophysiology Research: Sample Size, Standardization, Transparency, and Preregistration. *Biofeedback*, *48*(1), 2–6. https://doi.org/10.5298/1081-5937-48.1.2

Leach, S. C., Morales, S., Bowers, M. E., Buzzell, G. A., Debnath, R., Beall, D., & Fox, N. A. (2020). Adjusting ADJUST: Optimizing the ADJUST algorithm for pediatric data using geodesic nets. *Psychophysiology*, *57*(8), e13566. https://doi.org/10.1111/psyp.13566

Levin, A. R., Méndez Leal, A. S., Gabard-Durnam, L. J., & O'Leary, H. M. (2018). BEAPP: The Batch Electroencephalography Automated Processing Platform. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00513

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*(1), 358. https://doi.org/10.1038/s41467-019-14203-0

Lightman, B. V. (Ed.). (2016). *A companion to the history of science*. John Wiley & Sons.

Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., Barnes, G., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W., & Friston, K. (2011, March 6). *EEG and MEG Data Analysis in SPM8* [Research Article]. Computational Intelligence and Neuroscience; Hindawi. https://doi.org/10.1155/2011/852961

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2020). *Standardized Measurement Error: A Universal Measure of Data Quality for Averaged Event-Related Potentials (v20b)* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/dwm64

Marcus, E. (2016). A STAR Is Born. *Cell*, *166*(5), 1059–1060. https://doi.org/10.1016/j.cell.2016.08.021

Milham, M. P., Craddock, R. C., Son, J. J., Fleischmann, M., Clucas, J., Xu, H., Koo, B., Krishnakumar, A., Biswal, B. B., Castellanos, F. X., Colcombe, S., Di Martino, A., Zuo, X.-N., & Klein, A. (2018). Assessment of the impact of shared brain imaging data on the scientific literature. *Nature Communications*, *9*(1), 2818. https://doi.org/10.1038/s41467-018-04976-1

Miltner, W., Braun, C., Johnson, R., Simpson, G. V., & Ruchkin, D. S. (1994). A test of brain electrical source analysis (BESA): A simulation study. *Electroencephalography and Clinical Neurophysiology*, *91*(4), 295–310. https://doi.org/10.1016/0013-4694(94)90193-7

Moctezuma, L. A., & Molinas, M. (2020). Subject Identification from Low-Density EEG-Recordings of Resting-States: A Study of Feature Extraction and Classification. In K.

Arai & R. Bhatia (Eds.), *Advances in Information and Communication* (pp. 830–846). Springer International Publishing. https://doi.org/10.1007/978-3-030-12385-7_57

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229–240. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

Mourik, T. van, Snoek, L., Knapen, T., & Norris, D. G. (2018). Porcupine: A visual pipeline tool for neuroimaging analysis. *PLOS Computational Biology*, *14*(5), e1006064. https://doi.org/10.1371/journal.pcbi.1006064

National Academies of Sciences, E. (2019). *Reproducibility and Replicability in Science*. https://doi.org/10.17226/25303

Nave, K., Hannon, E., & Snyder, J. S. (2018). *Replication and Extension of Nozaradan, Peretz, Missal and Mouraux (2011)*. https://doi.org/None

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Niso, G., Gorgolewski, K. J., Bock, E., Brooks, T. L., Flandin, G., Gramfort, A., Henson, R. N., Jas, M., Litvak, V., T. Moreau, J., Oostenveld, R., Schoffelen, J.-M., Tadel, F., Wexler, J.,

& Baillet, S. (2018). MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Scientific Data*, *5*(1), 180110. https://doi.org/10.1038/sdata.2018.110

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2010, December 23). *FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data* [Research Article]. Computational Intelligence and Neuroscience; Hindawi. https://doi.org/10.1155/2011/156869

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Paul, M., Govaart, G., & Schettino, A. (2020). *Preregistration: A Solution to Undisclosed Analytic Flexibility in ERP Research*. PsyArXiv. https://doi.org/10.31234/osf.io/4tgve

Pavlov, Y. G., Adamian, N., Appelhoff, S., Arvaneh, M., Christopher Benwell, P. D., Beste, C., Bland, A., Bradford, D. E., Bublatzky, F., Busch, N., Clayson, P. E., Cruse, D., Czeszumski, A., Almenberg, A. D., Dumas, G., Ehinger, B. V., Ganis, G., He, X.,

Hinojosa, J. A., … Mushtaq, F. (2020). *#EEGManyLabs: Investigating the Replicability of Influential EEG Experiments*. PsyArXiv. https://doi.org/10.31234/osf.io/528nr

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., & Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, *6*(1), 103. https://doi.org/10.1038/s41597-019-0104-8

Peyk, P., DeCesarei, A., & Junghöfer, M. (2011). Electro Magneto Encephalograhy Software: Overview and integration with other EEG/MEG toolboxes. *Computational Intelligence and Neuroscience*, *2011*, Article ID 861705.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

Ram, K., Boettiger, C., Chamberlain, S., Ross, N., Salmon, M., & Butland, S. (2018). A community of practice around peer review for long-term research software sustainability. *Computing in Science & Engineering*, *21*(2), 59-65.

Regan, D. (1989). *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*. Elsevier.

Rule A, Birmingham A, Zuniga C, Altintas I, Huang S-C, Knight R, et al. (2019) Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. PLoS Comput Biol 15(7): e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Sandre, A., Banica, I., Riesel, A., Flake, J., Klawohn, J., & Weinberg, A. (2020). Comparing the effects of different methodological decisions on the error-related negativity and its association with behaviour and gender. *International Journal of Psychophysiology*, *156*, 18–39. https://doi.org/10.1016/j.ijpsycho.2020.06.016

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schwab, M., Karrenbach, N., & Claerbout, J. (2000). Making scientific computations reproducible. *Computing in Science Engineering*, *2*(6), 61–67. https://doi.org/10.1109/5992.881708

Shade, A., & Teal, T. K. (2015). Computing workflows for biologists: a roadmap. PLoS biology, 13(11), e1002303.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2019). *Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications* (SSRN Scholarly Paper ID 2694998). Social Science Research Network. https://doi.org/10.2139/ssrn.2694998

Smith, E. E., Tenke, C. E., Deldin, P. J., Trivedi, M. H., Weissman, M. M., Auerbach, R. P., Bruder, G. E., Pizzagalli, D. A., & Kayser, J. (2020). Frontal theta and posterior alpha in resting EEG: A critical examination of convergent and discriminant validity. *Psychophysiology*, *57*(2), e13483. https://doi.org/10.1111/psyp.13483

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sutton, B. P., Goh, J., Hebrank, A., Welsh, R. C., Chee, M. W. L., & Park, D. C. (2008). Investigation and validation of intersite fMRI studies using the same imaging hardware. *Journal of Magnetic Resonance Imaging : JMRI*, *28*(1), 21–28. https://doi.org/10.1002/jmri.21419

Swerdlow, N. R., Sprock, J., Light, G. A., Cadenhead, K., Calkins, M. E., Dobie, D. J., Freedman, R., Green, M. F., Greenwood, T. A., Gur, R. E., Mintz, J., Olincy, A., Nuechterlein, K. H., Radant, A. D., Schork, N. J., Seidman, L. J., Siever, L. J., Silverman, J. M., Stone, W. S., … Braff, D. L. (2007). Multi-site studies of acoustic startle and

prepulse inhibition in humans: Initial experience and methodological considerations based on studies by the Consortium on the Genetics of Schizophrenia. *Schizophrenia Research*, *92*(1), 237–251. https://doi.org/10.1016/j.schres.2007.01.012

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., & Leahy, R. M. (2011, April 13). *Brainstorm: A User-Friendly Application for MEG/EEG Analysis* [Research Article]. Computational Intelligence and Neuroscience; Hindawi. https://doi.org/10.1155/2011/879716

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123–138. https://doi.org/10.1111/psyp.12629

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Whiteford, K. L., Baltzell, L. S., Cooper, J. K., Irsik, V. C., Irvine, A., Mesik, J., Nolan, T., Oakes, B., Reed, A., & Schrlau, A. E. (2019). *Association of musical training with auditory and speech neural coding and perception*. https://doi.org/10/gg3d2k

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. PLoS computational biology, 13(6), e1005510.

Wilson, G. (2019). Teaching Tech Together: How to Make Your Lessons Work and Build a Teaching Community around Them. CRC Press.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., … Imamizu, H. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLOS Biology*, *17*(4), e3000042. https://doi.org/10.1371/journal.pbio.3000042

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., Trivedi, M. H., Weissman, M. M., Shinohara, R. T., & Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, *39*(11), 4213–4227. https://doi.org/10.1002/hbm.24241

Zeng, N., Zuo, S., Zheng, G., Ou, Y., & Tong, T. (2020). Editorial: Artificial Intelligence for Medical Image Analysis of Neuroimaging Data. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00480

Figures

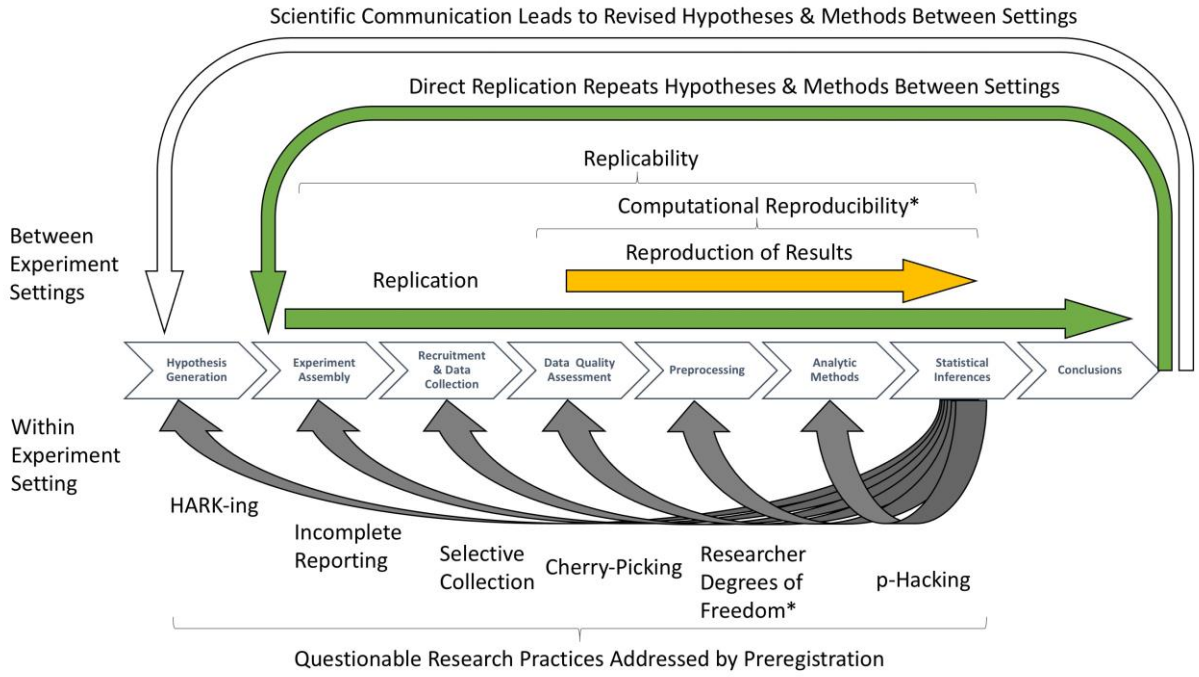

Figure 1. **Open science practices affect the research process at multiple levels**
The process of experimental research, involving steps ranging from hypothesis generation to drawing conclusions, is positively affected by various open science practices such as pre-registration and multi-laboratory studies. Direct replication requires sequential repetition of measurements and treatments. In a multisite study, identical measurements and treatments are carried out simultaneously between multiple similar experiment settings. Replicability in a multisite study thus supports the robustness of study outcomes. In this context, Computational Reproducibility addresses *Researcher Degrees of Freedom by constraining the influences of user defined parameters, code, and computing environment on analysis outcome. Likewise, preregistration precludes questionable research practices such as HARKing (hypothesizing after the results are known) by eliminating the possibility of outcome-dependent decision making.

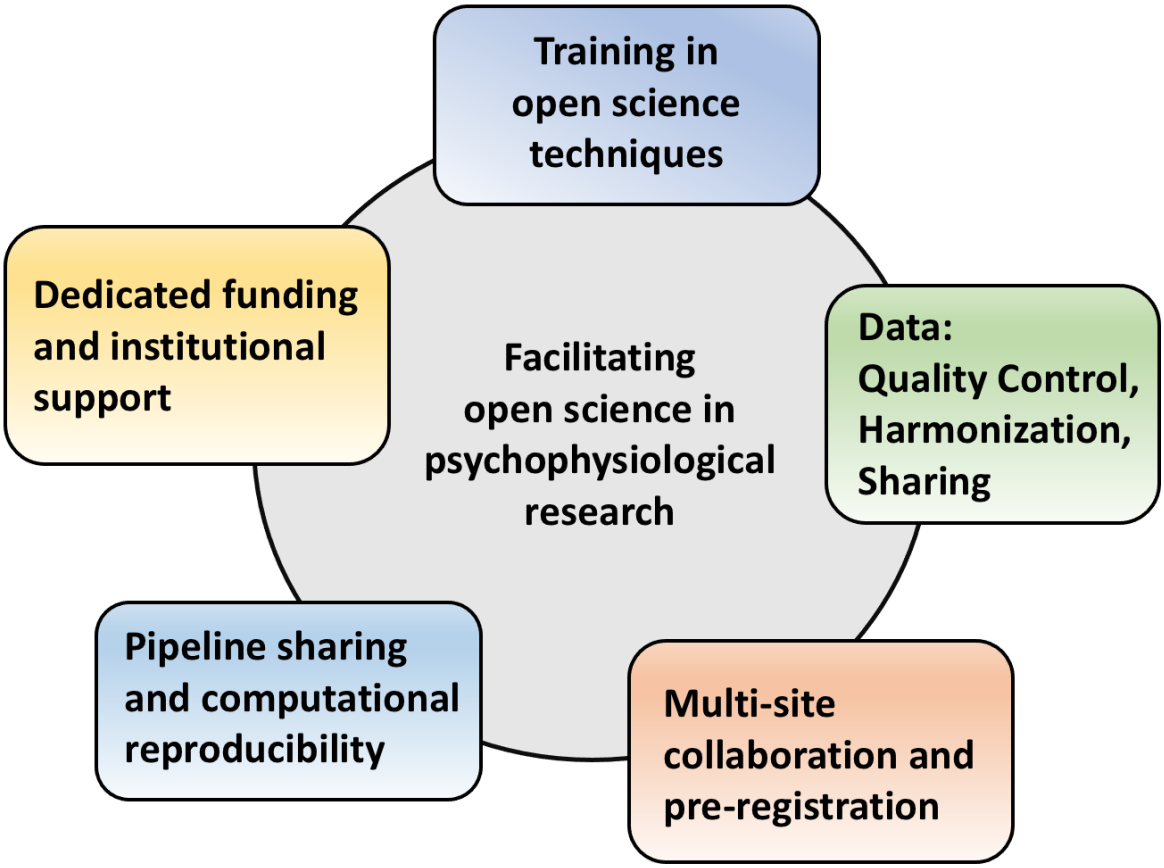

Figure 2. **Areas of opportunity for open science in psychophysiological research.** Advancing the five areas shown holds promise for expanding open science practices in psychophysiology. Integrating widely discussed open science practices that focus on the research process itself (pre-registration, data, sharing, pipeline sharing, etc.) with practices in training, funding, and collaboration may also address extant inequities in the access to the research process, including gender-related, geographical, racial, and economic inequities.